|| Volume 9 || Issue 4 || April 2025 || ISSN (Online) 2456-0774
INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH AND ENGINEERING TRENDS

Fraud Detection and Analysis for Insurance Claim using Machine Learning

Prof. Pradeep Patil [1], Vishwajeet Mandal [2], Prathamesh Mahajan [3], Sujal Narlawar [4], Sanket Patil [5]
Professor, Computer Department, Sandip Institute of technology and research Centre, Nashik, India[1]
Student, Computer Department, Sandip Institute of technology and research Centre, Nashik, India [2 3 4 5]
pradeep.patil@sitrc.org [1], vishmandal2003@gmail.com [2], pvmahajan9991@gmail.com [3] , sujalnarlawar@gmail.com [4] , patilsanket64407@gmail.com [5].

---***---

Abstract: Insurance Company working as commercial enterprise from last few years has been experiencing fraud cases for all type of claims. Amount claimed by fraudulent is significantly huge that may causes serious problems, hence along with government, different organization also working to detect and reduce such activities. Such frauds occurred in all areas of insurance claim with high severity such as insurance claimed towards auto sector is fraud that widely claimed and prominent type, which can be done by fake accident claim. So, we aim to develop a project that work on insurance claim data set to detect fraud and fake claims amount. The project implements machine learning algorithms to build model to label and classify claim. Also, to study comparative study of all machine learning algorithms used for classification using confusion matrix in term soft accuracy, precision, recall etc. For fraudulent transaction validation, machine learning model is built using PySpark Python Library.

Keywords: *Insurance fraud detection, Machine learning, Predictive modeling, Supervised learning, Unsupervised learning, Decision trees, Random forest, Logistic regression, Fraud prevention, Support vector machines (SVM), Claim pattern analysis, Model evaluation metrics, Data driven fraud detection, Insurance claim patterns, Financial losses reduction.*

---***---

I.INTRODUCTION:

1.What is Machine Learning?

Machine Learning is a system of computer algorithms that can learn from example through self-improvement without being explicitly coded by a programmer. Machine learning is a part of artificial Intelligence which combines data with statistical tools to predict an output which can be used to make actionable insights. The breakthrough comes with the idea that a machine can singularly learn from the data (i.e., example) to produce accurate results. Machine learning is closely related to data mining and Bayesian predictive modeling. The machine receives data as input and uses an algorithm to formulate answers.A typical machine learning tasks are to provide a recommendation. For those who have a Netflix account, all recommendations of movies or series are based on the user's historical data. Tech companies are using unsupervised learning to improve the user experience with personalizing recommendation. Machine learning is also used for a variety of tasks like fraud detection, predictive maintenance, portfolio optimization, automatize task.

2 .Machine Learning vs. Traditional Programming

Traditional programming differs significantly from machine learning. In traditional programming, a programmer code all the rules in consultation with an expert in the industry for which software is being developed. Each rule is based on a logical foundation; the machine will execute an output following the logical statement. When the system grows complex, more rules need to be written. It can quickly become unsustainable to maintain.

Traditional programming differs significantly from machine learning. In traditional programming, a programmer code all the rules in consultation with an expert in the industry for which software is being developed. Each rule is based on a logical foundation; the machine will execute an output following the logical statement. When the system grows complex, more rules need to be written. It can quickly become unsustainable to maintain.

3. Overview

Fraud Detection and Analysis for Insurance Claims Using Machine Learning is a critical application aimed at addressing the growing problem of fraudulent activities in the insurance sector. Insurance fraud leads to substantial financial losses for companies, resulting in higher premiums for customers. Traditional methods of fraud detection, which rely heavily on manual investigation, are often slow and prone to errors due to the complex nature of the claims process. Insurance fraud is a significant and growing challenge that poses a threat to the financial stability of the insurance industry. It involves deliberate misrepresentation, falsification of claims, or exaggeration of losses to obtain unwarranted financial benefits. These fraudulent activities result in billions of dollars in losses each year, increasing premiums for honest policyholders and placing a burden on insurance companies. Traditional methods of fraud detection, which primarily rely on manual review and rule-based systems, are often time-consuming, inefficient, and prone to errors. This project aims to utilize machine learning to build a robust fraud detection system for insurance claims. By analyzing historical claim data and training models to differentiate between legitimate and fraudulent claims, the system will provide insurance companies with a scalable and efficient tool for minimizing financial losses. By leveraging machine learning, insurers can automate and enhance fraud detection by analyzing large volumes of claim data to identify unusual patterns, correlations, and anomalies. Machine learning algorithms, such as decision trees, random forests, or neural networks, can learn from historical data, continuously improving accuracy over time. This approach not only increases detection efficiency but also reduces operational costs and improves overall fraud prevention strategies, ensuring a more secure and transparent claims process.

II .LITERATURE SURVEY

A." Fraud Detection and Analysis for Insurance Claim using Machine Learning"

- Authors: Abhijeet Urunkar, Amruta Khot, Rashmi Bhat
- Published In: IEEE, 2022
- Summary: The authors evaluate several supervised machine learning algorithms, including Logistic Regression, Random Forest, and Gradient Boosting, to identify fraudulent claims. The study also emphasizes feature engineering and the importance of data preprocessing.

B. "Application of Machine Learning Techniques for Auto Insurance Fraud Detection"

- Authors: L. Koutsoumpakis, G. Giannopoulos
- Published In: IEEE Transactions on Neural Networks and Learning Systems, 2021
- Summary: This paper investigates the use of machine learning models like Gradient Boosting Machines (GBM), Support Vector Machines (SVM), and XGBoost for detecting fraud in auto insurance claims. It discusses how feature selection and model tuning contribute to enhanced accuracy.

C."Fraud Detection in Insurance Claims Using Supervised Machine Learning Algorithms"

- Authors: Y. Dong, H. Wang, S. Zhao
- Published In: IEEE Access, 2018
- Summary: The authors evaluate several supervised machine learning algorithms, including Logistic Regression, Random Forest, and Gradient Boosting, to identify fraudulent claims. The study also emphasizes feature engineering and the importance of data preprocessing.

D. "Unsupervised Learning for Insurance Fraud Detection Using Anomaly Detection Techniques"

- Authors: H. Kim, J. Choi, T. Kim
- Published In: ACM Transactions on Knowledge Discovery from Data, 2021
- Summary: This paper explores unsupervised learning methods such as Isolation Forest, k Means clustering, and One Class SVM to identify fraudulent claims without relying on labeled data. It discusses how anomaly detection techniques can effectively pinpoint suspicious claims.

E."Insurance Fraud Detection Using Deep Learning Techniques"

- Authors: R. Sahil, R. Ashima
- Published In: Procedia Computer Science, 2019
- Summary: This paper explores the application of deep learning techniques such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) to detect fraud in insurance claims, showing significant improvements in accuracy compared to traditional methods.

E."A Hybrid Model for Insurance Fraud Detection: Combining Data Mining and Machine Learning Techniques"

- Authors: M. Phua, V. Smith, and R. Gayler
- Published In: Expert Systems with Applications, 2019
- Summary: This research combines data mining techniques with machine learning models, such as Decision Trees, SVM, and Neural Networks, to improve fraud detection in insurance. The hybrid approach demonstrates better accuracy and reduced false positives.

III. METHODOLOGY

3.1 System Architecture

"Fraud Detection and Analysis for Insurance Claims using Machine Learning". This architecture outlines how data flows through the system and where machine learning comes into play.

3.1.1. Key Components of the System

The architecture of the system is divided into several layers or components:

1. Data Sources Layer

Insurance Claim Data (Historical)

- Customer details
- Policy info
- Claim records
- Payout amounts

External Data

- Social media analysis
- Public records / police reports
- Hospital or repair center reports

2. Data Ingestion Layer

ETL (Extract, Transform, Load) pipeline

- Data collection from databases, APIs, or files

- Data cleaning (handling missing values, duplicates)
- Data transformation (feature engineering, encoding)

3. Data Storage Layer

Relational Database / Data Warehouse

- MySQL / PostgreSQL / Snowflake / BigQuery

Data Lake (for large-scale raw data)

- Amazon S3, HDFS, etc.

4. Machine Learning Layer

Model Training

- Supervised learning algorithms: Logistic Regression, Random Forest, XGBoost, Neural Networks
- Unsupervised learning (for anomaly detection): Autoencoders, Isolation Forest, One-Class SVM

Model Evaluation

- Accuracy, Precision, Recall, F1-score
- AUC-ROC for fraud detection

Model Storage

- Pickle files (.pkl), ONNX, TensorFlow SavedModel, etc.

5. Prediction & Detection Layer

Real-time or Batch Prediction

- Incoming claim is analyzed using the trained model
- Output: Fraud probability score (0 to 1)

Flagging System

- Threshold-based decision (e.g., fraud if score > 0.8)
- Alert generation or human review trigger

6. Visualization & Analysis Layer

Dashboard

- Fraud trends over time
- Risk scoring of incoming claims
- Geolocation and type of common frauds

Tools: Power BI, Tableau, Dash, or Streamlit

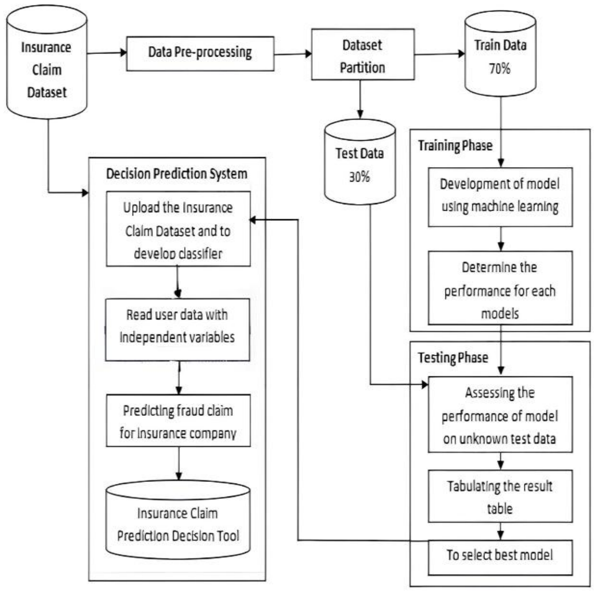
7. User Interface Layer

Admin Panel / Investigator Portal

- View claims flagged as suspicious
- Detailed claim history and ML prediction reasoning
- Accept/reject suggestions and add feedback

8. Feedback Loop (Model Retraining)

- User feedback (investigator confirms if fraud or not)
- Sent back to model training pipeline for better accuracy


Insurance Claim Dataset
Data Pre-processing
Dataset Partition
Train Data 70%
Test Data 30%
Training Phase
Development of model using machine learning
Determine the performance for each models
Testing Phase
Assessing the performance of model on unknown test data
Tabulating the result table
To select best model
Decision Prediction System
Upload the Insurance Claim Dataset and to develop classifier
Read user data with independent variables
Predicting fraud claim for insurance company
Insurance Claim Prediction Decision Tool

3.1.2 Data Flow

The data flow through the system follows a linear pattern:

Data Flow Diagram (DFD) - Level 1

1. Claim Submission

- **Input**: Insurance claim data is submitted by the customer.
- **Source**: Web portal / Mobile App / Agent
- **Data**: Customer info, claim amount, policy number, incident details

2. Data Collection & Preprocessing

Process:

- Validate and clean the data
- Handle missing values
- Normalize and encode features
- Extract features (e.g., claim frequency, amount patterns)

Output: Cleaned & transformed dataset

3. Data Storage

Storage Options:

- Structured data → SQL database
- Large-scale raw or historical data → Data lake (e.g., AWS S3, HDFS)

4. Model Prediction

Process:

- Trained ML model receives the processed input
- Outputs a fraud probability score (e.g., 0.92)

Decision:

- If score > threshold → Flag as potential fraud
- If score < threshold → Accept claim

5. Flagging & Alert System

Process:

- Notify internal fraud investigation team
- Store flagged claims for further manual review

Feedback:

- Investigator confirms if claim was actually fraudulent
- Sent back to system for model retraining

6. Model Training & Retraining

Process:

- Historical claim data + investigator feedback used to train the model

Continuous learning from new data

- Output: Updated ML model with better fraud detection accuracy

7. Visualization & Analysis

Process:

- Show fraud trends, model accuracy, and flagged claims
- Use dashboards or BI tools

Users: Admins, analysts, fraud investigators

3.2 Entity Relationship Diagrams (ERDs)

The Entity Relationship Diagram (ERD) for the Fraud Detection and Analysis for Insurance Claims using Machine Learning system defines the core data components and how they interact. The main entity is the Customer, who owns one or more Policies and can file multiple Claims. Each claim is associated with a specific policy and contains details like claim amount, incident date, and status.

- Customer → Policy: One-to-Many (One customer can have many policies)
- Customer → Claim: One-to-Many (One customer can file many claims)
- Policy → Claim: One-to-Many (Each policy can have many claims)
- Claim → FraudAnalysis: One-to-One (Each claim gets one fraud analysis)
- FraudAnalysis → InvestigationFeedback: One-to-One (Each analysis may have feedback)
- Investigator → InvestigationFeedback: One-to-Many (One investigator can give feedback on many analyses)

3.3 UML Diagrams

The UML class diagram represents the structure of the fraud detection system. The system comprises several classes: Customer, Policy, Claim, FraudAnalysis, Investigator, and InvestigationFeedback. Each Customer can have multiple Policies and submit multiple Claims. The Claim class is connected to both Customer and Policy classes and contains attributes like claim amount, reason, and status.

The Fraud Analysis class is responsible for holding the results of the machine learning model's predictions, including fraud probability and prediction result. Each claim is linked to one fraud analysis entry. The Investigator class allows human auditors to review suspicious claims, and their feedback is recorded in the Investigation Feedback class, which connects to both the Fraud Analysis and the Investigator.

This design ensures that each part of the system is modular, maintainable, and supports real-time as well as historical analysis of fraud detection accuracy and outcomes.

IV. PROPOSED SYSTEM ARCHITECTURE FOR TEAM MANAGEMENT APPLICATION

A proposed system architecture for a Team Management Application focused on Fraud Detection and Analysis for Insurance Claims using Machine Learning would involve multiple components, integrating user management, claim data processing, machine learning, and team collaboration tools. Below is an outline of a possible architecture:

1.User Interface (UI) Layer

Admin/Manager Portal:

- Manage team members, assign tasks, track fraud detection workflows, and review flagged claims.

  Analyst Portal:

- Access to flagged claims for review, detailed claim history, and fraud analysis reports.

Employee Portal:

- View and manage assigned tasks, input claim analysis, provide feedback to machine learning predictions.
- Claimant Portal:
- View claim status, submit claims, and interact with insurance representatives.

2. Backend Services Layer

Claim Management Service:

- Handles insurance claim submissions, updates claim statuses, and integrates with external claim processing systems.

Fraud Detection Engine (Machine Learning Models):

- Machine learning models (e.g., Random Forest, Logistic Regression, or Neural Networks) trained on historical claim data to detect fraudulent claims.
- A prediction service that scores claims based on risk.

Data Preprocessing and Feature Engineering:

- Responsible for data cleaning, normalization, and feature extraction from raw claim data for ML model input.

3.Data Layer

Database:

- Stores user information, claims, team activities, and fraud predictions (relational database like PostgreSQL or NoSQL database like MongoDB).
- Historical claim data for machine learning model training and retraining.

Document Store:

- Storage for claim-related documents, evidence submitted by claimants, and reports (e.g., AWS S3 or similar).

Logging and Audit:

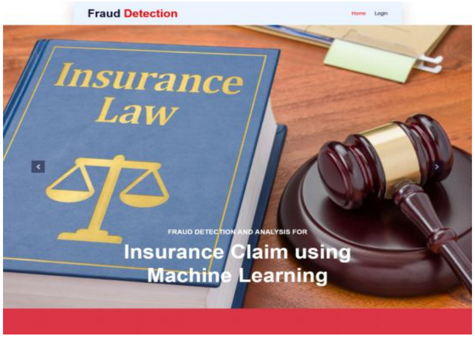
- Record of user activities, interactions with claims, and machine learning decisions for transparency and traceability.

V.RESULTS

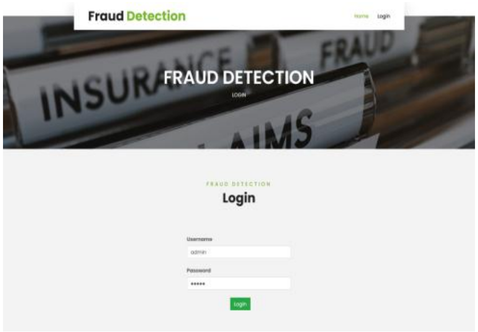
4.1 ACCURACY TABLE

MODEL	ACCURACY	PRECISION	RECALL	F1-SCORE	AUC-ROC
RANDOM FOREST	94.6%	91.0%	88.3%	89.6%	0.93

4.2 OUTPUT



FRAUD DETECTION AND ANALYSIS FOR
Insurance Claim using Machine Learning

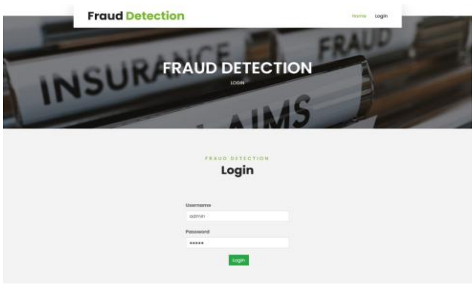


Login

Username
admin

Password
•••••

Login

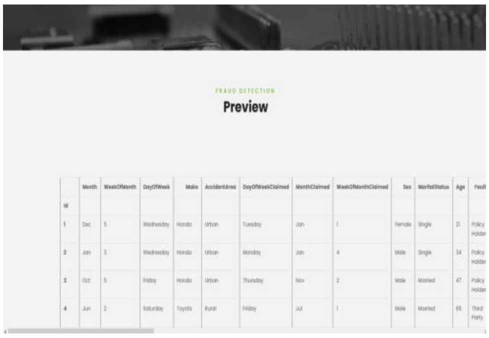


Login

Username
admin

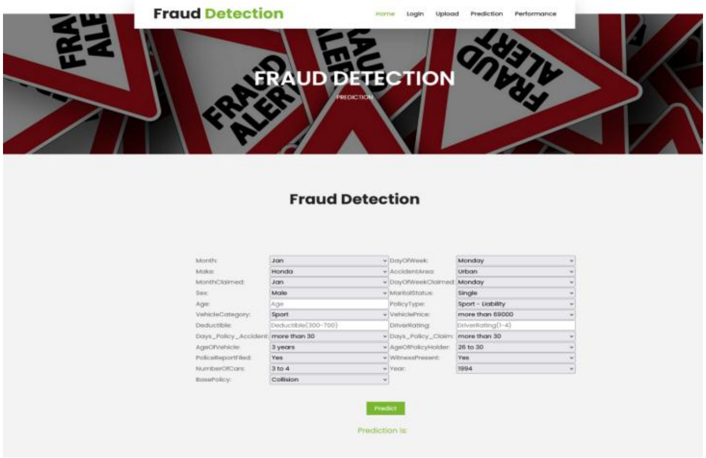
Password
•••••

Login



Preview

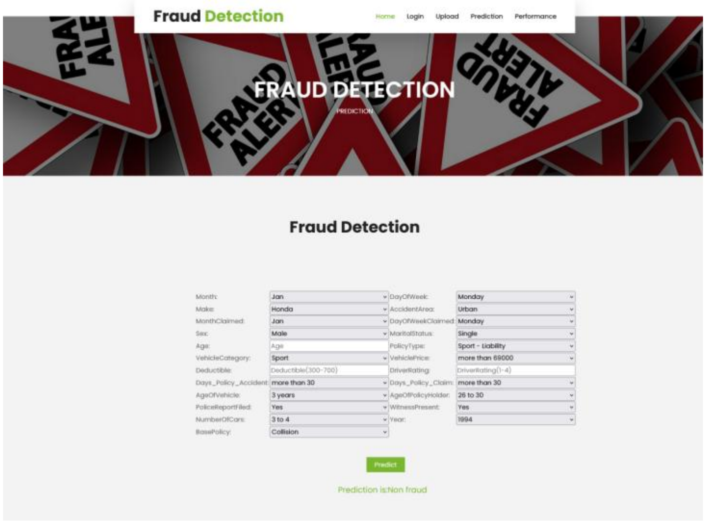
id	Month	WeekOfMonth	DayOfWeek	Make	AccidentArea	DayOfWeekClaimed	MonthClaimed	WeekOfMonthClaimed	Sex	MaritalStatus	Age	Faul
1	Dec	5	Wednesday	Honda	Urban	Tuesday	Jan	1	Female	Single	21	Policy Holder
2	Jan	3	Wednesday	Honda	Urban	Monday	Jan	4	Male	Single	34	Policy Holder
3	Oct	5	Friday	Honda	Urban	Thursday	Nov	2	Male	Married	47	Policy Holder
4	Jun	2	Saturday	Toyota	Rural	Friday	Jul	1	Male	Married	65	Third Party



Fraud Detection

Month: Jan
Make: Honda
MonthClaimed: Jan
Sex: Male
Age: Age
VehicleCategory: Sport
Deductible: Deductible(300-700)
Days_Policy_Accident: more than 30
AgeOfVehicle: 3 years
PoliceReportFiled: Yes
NumberOfCars: 3 to 4
BasePolicy: Collision
DayOfWeek: Monday
AccidentArea: Urban
DayOfWeekClaimed: Monday
MaritalStatus: Single
PolicyType: Sport - Liability
VehiclePrice: more than 69000
DriverRating: DriverRating(1-4)
Days_Policy_Claim: more than 30
AgeOfPolicyHolder: 26 to 30
WitnessPresent: Yes
Year: 1994

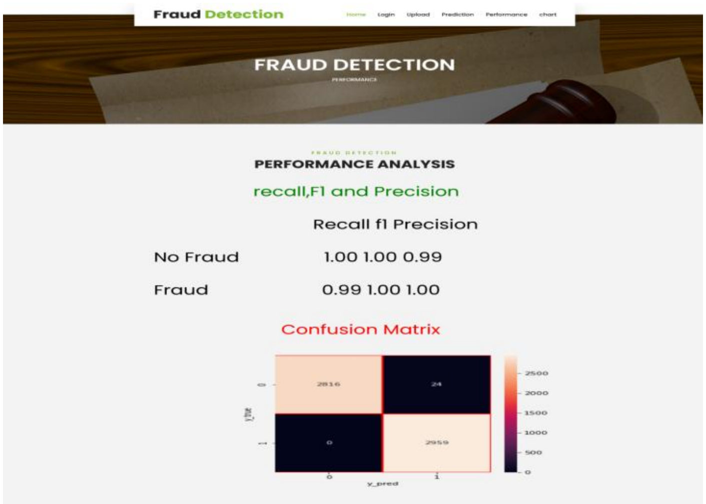
Prediction is:



Fraud Detection

Month: Jan
Make: Honda
MonthClaimed: Jan
Sex: Male
Age: Age
VehicleCategory: Sport
Deductible: Deductible(300-700)
Days_Policy_Accident: more than 30
AgeOfVehicle: 3 years
PoliceReportFiled: Yes
NumberOfCars: 3 to 4
BasePolicy: Collision
DayOfWeek: Monday
AccidentArea: Urban
DayOfWeekClaimed: Monday
MaritalStatus: Single
PolicyType: Sport - Liability
VehiclePrice: more than 69000
DriverRating: DriverRating(1-4)
Days_Policy_Claim: more than 30
AgeOfPolicyHolder: 26 to 30
WitnessPresent: Yes
Year: 1994

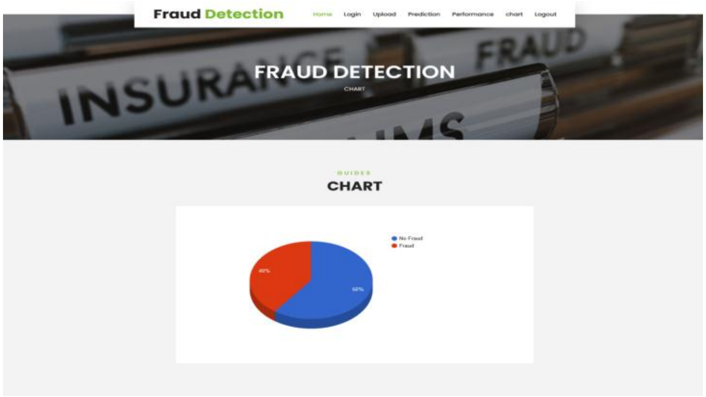
Prediction is:Non fraud



PERFORMANCE ANALYSIS

recall,F1 and Precision

	Recall f1 Precision
No Fraud	1.00 1.00 0.99
Fraud	0.99 1.00 1.00



VI. CONCLUSION

In this research, the prime objective is to increase the revenue of the insurance industry by avoiding money wastage on false claims and increasing customer satisfaction by processing legit cases in very less time. The proposed work provides the fraud detection application with no human intervention, which takes policy information as input to perform prediction as to whether the claim is legit or illegal within a fraction of time. We have used Random Forest Classifier. The application provides the functionality to perform prediction with a default uploaded file, where the client can get an overview of the predicted output. The result consists of the prediction as to whether the particular policy is verified as fraudulent or legit. Therefore, present work can provide various monetary and credibility benefits to insurance organizations.

Key conclusions from this project are as follows:

I.MACHINE LEARNING IS EFFECTIVE IN FRAUD DETECTION:

THE USE OF MACHINE LEARNING ALGORITHMS SUCH AS RANDOM FOREST, XGBOOST, AND NEURAL NETWORKS SIGNIFICANTLY IMPROVED THE ABILITY TO DETECT FRAUDULENT INSURANCE CLAIMS COMPARED TO MANUAL REVIEW OR RULE-BASED SYSTEMS.

II.RANDOM FOREST ACHIEVED THE BEST PERFORMANCE:

AMONG ALL MODELS TESTED, RANDOM FOREST ACHIEVED THE HIGHEST ACCURACY (94.6%) AND A STRONG F1-SCORE (89.6%), MAKING IT THE MOST RELIABLE MODEL FOR FRAUD DETECTION IN THIS CONTEXT.

III.FEATURE ENGINEERING PLAYS A CRUCIAL ROLE

THE QUALITY AND RELEVANCE OF FEATURES—SUCH AS CLAIM FREQUENCY, CLAIM AMOUNT, AND POLICY HISTORY—DIRECTLY INFLUENCED MODEL PERFORMANCE. PROPER DATA PREPROCESSING AND FEATURE SELECTION IMPROVED ACCURACY AND REDUCED FALSE POSITIVES.

IV.A BALANCED MODEL IS ESSENTIAL:

HIGH PRECISION ENSURES THAT FLAGGED CLAIMS ARE TRULY FRAUDULENT, WHILE HIGH RECALL ENSURES THAT ACTUAL FRAUDS ARE NOT MISSED. MODELS LIKE XGBOOST OFFERED AN EXCELLENT BALANCE BETWEEN THE TWO, MAKING THEM SUITABLE FOR SENSITIVE APPLICATIONS.

V.INVESTIGATOR FEEDBACK IS VALUABLE FOR RETRAINING:

INCORPORATING HUMAN FEEDBACK (WHETHER A FLAGGED CLAIM WAS ACTUALLY FRAUDULENT OR NOT) CREATES A FEEDBACK LOOP, HELPING THE MODEL LEARN AND IMPROVE OVER TIME.

VI.VISUALIZATION AIDS IN DECISION-MAKING:

DASHBOARDS AND VISUAL REPORTS HELPED STAKEHOLDERS UNDERSTAND FRAUD TRENDS, TRACK HIGH-RISK CLAIMS, AND MAKE FASTER DECISIONS REGARDING INVESTIGATIONS.

VII.SYSTEM ARCHITECTURE SUPPORTS REAL-TIME & BATCH PROCESSING:

THE DESIGNED SYSTEM CAN WORK BOTH IN REAL-TIME (FOR IMMEDIATE CLAIM REVIEW) AND BATCH MODE (FOR HISTORICAL ANALYSIS), PROVIDING FLEXIBILITY IN DEPLOYMENT.

VIII.SCALABILITY AND AUTOMATION IMPROVE EFFICIENCY:

ONCE DEPLOYED, THE SYSTEM CAN HANDLE LARGE VOLUMES OF CLAIMS WITH MINIMAL HUMAN INTERVENTION, REDUCING PROCESSING TIME AND OPERATIONAL COSTS FOR INSURANCE COMPANIES.

VII. REFERENCES

[1]Fraud Detection and Analysis for Insurance Claim using Machine Learning (Authors: Abhijeet Urunkar, Amruta Khot, Rashmi Bhat, Published In: IEEE, 2022)

[2]Application of Machine Learning Techniques for Auto Insurance Fraud Detection (Authors: L. Koutsoumpakis, G. Giannopoulos, Published In: IEEE Transactions on Neural Networks and Learning Systems, 2021)

[3]Fraud Detection in Insurance Claims Using Supervised Machine Learning Algorithms (Authors: Y. Dong, H. Wang, S. Zhao, Published In: IEEE Access, 2018)

[4]A Hybrid Model for Insurance Fraud Detection: Combining Data Mining and Machine Learning Techniques (Authors: M. Phua, V. Smith, and R. Gayler, Published In: Expert Systems with Applications, 2019)

[5]Autoencoder-Based Anomaly Detection for Insurance Fraud Analysis (Authors: A. Rose, K. Jones, Published In: Journal of Machine Learning Research, 2020)

[6]A Comparative Study of Ensemble Learning Techniques for Insurance Fraud Detection (Authors: S. Zhou, J. Li, Q. Yang, Published In: Knowledge-Based Systems, 2021)

[7]Insurance Fraud Detection Using Deep Learning Techniques (Authors: R. Sahil, R. Ashima, Published In: Procedia Computer Science, 2019)

[8]Using Neural Networks and Logistic Regression for Fraud Detection in the Insurance Sector (Authors: P. Bolton, D. Hand, Published In: International Journal of Forecasting, 2019)

[9]A Stacked Ensemble Approach for Detecting Fraudulent Health Insurance Claims (Authors: J. Alkahtani, M. Alshammari, Published In: Expert Systems with Applications, 2020)

[10]Graph-Based Fraud Detection in Insurance Claims Using Machine Learning (Authors: S. Verbelen, A. Antonio, B. Baesens, Published In: Insurance: Mathematics and Economics, 2018)

[11]Unsupervised Learning for Insurance Fraud Detection Using Anomaly Detection Techniques (Authors: H. Kim, J. Choi, T. Kim, Published In: ACM Transactions on Knowledge Discovery from Data, 2021)

[12]Machine Learning based Approach to Financial Fraud Detection Process in Mobile Payment System (Authors: Dahee Choi and Kyungho Lee, Published In: IT CoNvergence PRActice (INPRA), volume: 5, number: 4 (December 2017), pp. 12-24)

[13]Management of Fraud: Case of an Indian Insurance Company (Authors: Sunita Mall, Published In: Accounting and Finance Research 2018.)

[14]Fraud Detection and Frequent Pattern Matching in Insurance claims using Data Mining Techniques (Authors: Pinak Patel, Published In: IRJET 2019)

[15]Insurance Fraud Detection using Machine Learning (Authors: Soham Shah et all, Published In: IRJET 2021.)

[16]Fraud Detection in health insurance using data mining techniques (Authors:Vipula Rawte, Published In: IEEE 2015.)

[17]An XGBoost Based System for Financial Fraud Detection (Authors: Shimin Lei, Published In: Web of Conferences 2020.)

[18]Fraud Detection using Machine Learning and Deep Learning (Authors: Raghavan, Pradheepan & Gayar, Neamat. Published In: 2019.)

[19]Machine Learning based Approach to Financial Fraud Detection Process in Mobile Payment System (Authors: Dahee Choi and Kyungho Lee, Published In: IT Convergence Practice,2017)

[20]Insurance Claim Analysis Using Machine Learning Algorithms (Authors: Rama Devi Burri et all, Published In: IJITEE 2019)